# STA 250F — Assignment #1 — Due in lecture on October 31 at 11:10am

*Late assignments will be accepted only with a valid medical or other excuse.*
*This assignment is to be done by each student individually.*
*There are four pages to this assignment.*

For both this assignment and the next you will look at data for the problem described below. This data was produced by a fairly realistic computer simulation of the situation described. Each of you has *different* data, for a situation that differs in detail from the situation for other students. You **must** analyse your own data, not someone else's. The conclusions you come to may be different from those of some other student, since your data, and also the underlying situation, will be different from theirs.

Your aim in this assignment should be to analyse the data as if this were a real problem. You should do whatever is necessary (within the limits of what we've covered in the course so far) in order to get good answers to whatever questions would be of interest if this were a real problem. To help you in this, certain specific questions are mentioned below, which you should certainly address, but other questions might occur to you as well. You might also consider the various techniques we have covered in the course so far, and decide which of them might be useful for this problem. Note that you can expect the data to show some variation due to chance. For this assignment, you will have to judge what is just chance informally, since we haven't covered how to do it formally yet.

## The problem: Analysis of an experiment on the effect of exercise and dietary restriction on the lifespan of mice

It is commonly thought that exercise and proper diet increase lifespan in humans. Some past experiments on animals have shown an increase in lifespan when the amount of food consumed is restricted. It's also possible that diet and exercise might interact — for instance, exercise might be good if you eat lots of food, but bad if your diet is more restricted. The effects might perhaps be non-linear — for instance, more food might lead to longer lifespans up to a point, but beyond that, more food might lead to shorter lifespans.

This assignment deals with an (imaginary) experiment designed to see what effect exercise and dietary restriction have on lifespan in mice.

Two hundred mice were used in the experiment. All mice were of the same variety, but this was not an inbred variety, so there will be some genetic variation among the mice. None of the mice were closely related. The mice were all raised to age 100 days in standard laboratory conditions. Mice are fully grown at this age.

The 100-day-old mice were randomly divided into ten groups, with 20 mice in each group (10 males and 10 females). The groups differed in whether or not the mice were allowed to exercise, and in how much food they were allowed to eat. Each mouse was housed in a separate cage (but near the others, so they wouldn't feel lonely). Some of the cages contained an exercise wheel in which the mice could run around, with a device that recorded how many rotations of the wheel occurred each day. The other cages did not have any equipment that facilitated exercise, and were too small for the mice to get much exercise just by running around. Each day, a controlled quantity of food was put in each mouse's cage. At the end of the day, any uneaten food was removed. The amount of food uneaten was weighed, allowing the amount of food the mouse ate to be found by subtracting this from the amount that was put in the cage.

Five of the ten groups were allowed exercise, the other five groups could not exercise. The five exercise groups differed in the amount of food they were allowed, which was 4, 5, 6, 8, or 10 grams per day. These amounts of food allowed are intended to range from the minimum amount that sustains life to as much as a mouse might want to eat. The five non-exercise groups differed in the same way, with the same amounts of food allowed.

Each mouse was weighed at the beginning of the experiment (when they were 100 days old), and after 100 days (when they were 200 days old). This might give some indication of whether the mice were becoming obese, and of whether they were healthy.

The experiment continued until all the mice had died of natural causes. A record was made of the age of each mouse when it died (ie, of its lifespan).

Note this is a *randomized experiment*, since the experimenters randomly divided the mice into groups, and then controlled how much food mice in each group were given, and whether or not they were allowed exercise. From experimental data such as this, we can hope to draw conclusions about whether or not allowing the mice to exercise or allowing them to eat lots of food actually *causes* an increase or decrease in lifespan.

In contrast, if we allowed all the mice to exercise if they liked, and to eat as much as they wanted, and then looked at how much they actually exercised, how much they actually ate, and how long they lived, we would have only *observational data*. Determining whether or not the amount of exercise or amount of food eaten actual had an effect on lifespan would then be difficult, because of the possibility of confounding influences.

## The data format

You can read in the data using File > Other Files > Import Special Text. The data you should use is in the file with the same name as your CQUEST account, located in the directory called assignment-data. *Be sure you read your data, not someone else's data.* If you are doing the assignment from home, you can download your data from the course web page. **Note:** The data is not available yet, but will be by October 10.

The data file consists of 200 lines, one for each mouse, with each line containing ten numbers, as follows:

| | | |
|---|---|---|
| 1 | ex-allowed | 0 if the mouse was not allowed to exercise, 1 if it was allowed to exercise |
| 2 | food-allowed | The amount of food per day the mouse was allowed to eat (grams) |
| 3 | sex | The mouse's sex: 0=male, 1=female |
| 4 | ex-amount | The average amount of exercise per day (rotations of wheel) |
| 5 | food-amount | The average amount of food eaten per day (grams) |
| 6 | weight100 | The mouse's weight at age 100 days (grams) |
| 7 | weight200 | The mouse's weight at age 200 days, "*" if the mouse died before then |
| 8 | lifespan | The age of the mouse at death (days) |
| 9 | ID | A numeric identifier for this mouse |
| 10 | allowed-max | Set to 1 if the mouse was allowed to exercise and also allowed the maximum amount of food; set to 0 otherwise |

The 'ex-amount' and 'food-amount' variables are averages over the first 100 days of the experiment. For mice that died within the first 100 days, these values are the averages from the start of the experiment until the mouse died. The 'ex-amount' variable is set to 0 for the mice that weren't allowed to exercise. The "*" value that 'weight200' has when the mouse died before age 200 days is interpreted by MINITAB as a "missing value". MINITAB will ignore units that have missing values for variables that are needed for what is being done at the moment.

The 'allowed-max' variable is determined by 'ex-allowed' and 'food-allowed', but is included to make it easier to "unstack" the data (as described below).

The names for the variables above ('ex-allowed', 'food-allowed', etc.) do not appear in the data file, but you should give the variables these names in MINITAB, by typing the name at the head of each column.

## Initial examination of the data

You should start by looking at the the distributions of the variables, using stem plots or histograms. Looking at these distributions for subsets of the mice may also be useful. You should also look at scatterplots for some pairs variables, such as 'food-amount' and 'food-allowed'.

One reason for looking at such plots is to identify data points that seem to be extreme, or that don't follow the relationship that the other points seem to follow. These might be data recording errors, or they might be due to exceptional events. You should try to determine whether such values are really erroneous, and if you decide that they are in error, you should set the value to "*", which is MINITAB's symbol for a "missing value". MINITAB will then ignore this datapoint. If you can't determine that an extreme value is the result of an error, or is due to an extraordinary event that you aren't interested in, then you shouldn't ignore it completely, but it might be advisable to see how much effect removing that value has on your later conclusions, and indicate that your conclusions are uncertain if they depend heavily on whether or not that value is included.

You should also consider whether the various variables have what look like normal distributions, or instead have distributions that are skewed, or perhaps bimodal. Does this or anything else you can see tell you anything about whether the experiment worked as intended?

## Looking at the mice that were allowed exercise and lots of food

If we look only at the 20 mice that were allowed to exercise and that were allowed the maximum amount of food (10 grams per day), what we have is effectively observational data. For these 20 mice, which were treated identically by the experimenter, we can see how lifespan is related to how much exercise the mice chose to do and how much food they chose to eat.

As an exercise in how the relationships seen in observational data may differ from the results of a randomized experiment, you should look at the data on these 20 mice, pretending for the moment that this is the only data you have. The 'allowed-max' variable is set to 1 for these mice, so that the data on them can easily be separated from the rest using the "unstack" facility in MINITAB.

To start, you can look at scatterplots of 'lifespan' versus 'ex-amount' and 'lifespan' versus 'food-amount'. You could also do regressions of 'lifespan' on each of these two variables. Looking at residual plots for these regressions may reveal additional features of the data. You could also produce subsets of the data (using "unstack" again), and do separate regressions on these subsets. The 'weight100' and 'weight200' variables may provide additional information on what could be going on.

You should state your tentative conclusions from this observational data. As we've discussed in lecture, these conclusions could be wrong because of confounding influences, perhaps involving variables that haven't been measured. This is why experimental data is a better way to find cause and effect relationships, if it can be obtained.

## Looking at the full experimental data set

You should now go on to look at all 200 mice from the experiment. The first thing to do is to look at how lifespan differed between the ten groups of mice. You could also look at how lifespan differed according to whether or not the mice were allowed exercise (considering all amounts of food allowed), or according to how much food they were allowed (with or without exercise being allowed). Since these variables were set by the experimenter, using randomization, these relationships can be interpreted in terms of cause and effect, without any possible confusion from unknown confounding influences. However, the relationships seen will still be subject to random variation.

It may also be useful to do a regression of 'lifespan' on 'food-allowed', but you should check whether the relationship is actually linear. You might also split up the data (eg, by 'sex' or by 'ex-allowed'), and do several separate regressions. You can also consider doing regressions of 'lifespan' on 'food-amount', or of 'lifespan' on 'ex-amount'. Looking at various other possible relationships might also be useful — for example, the 'weight100' and 'weight200' variables might help in understanding why the relationships are as they are. Your aim should be to find out not only what are the effects of diet and exercise on lifespan, but if possible, *why* the effects are as they are, and whether these effects are the same for all mice.

## Organization of your report

You should hand in a report that describes how you went about your analysis, and what your conclusions were. These conclusions should be supported by a moderate amount of MINITAB output with relevant plots and statistics. One way to conveniently include MINITAB output in your report is to create it by editing your session window. Or you could save your session window and then edit it with your favourite text editor.

Your report *must* be organized into sections with the following titles:

1. **Initial examination of the data**. Here you will report on any erroneous observations that you deleted (including why you think they are erroneous), on general features of the distributions, and on any conclusions you came to about how well the experiment worked.

2. **Observations on mice allowed exercise and lots of food**. Here you should report what relationships you found looking at only the 20 mice who were allowed to exercise, and were allowed the largest amount of food. Since this is effectively observational data, these relationships may be affected by confounding influences.

3. **Analysis of the experimental data**. This is the main section, in which you examine what seem to be the effects of diet and exercise on lifespan, and if possible, what seems to be responsible for these effects. Comment as well as you can on which features of the data you believe are due to real effects, and which might just be the result of random variation.

4. **Comparison of results from observational and experimental data**. Here you should compare what you saw in Sections 2 and 3 above. If the relationship you saw looking at all the experimental data is different from what seemed to be the case when looking only at the observational data, comment on what confounding influences might be responsible for this.

5. **Conclusions**. Here you should summarize the main results of your analysis (of the whole experimental data set). Your summary should address the question that the experiment was designed to answer — how diet and exercise affects lifespan. For instance, you should not just say "X and Y were positively correlated", but say also what this means as far as the effects on lifespan. You can also comment on anything else interesting that you found out from the data.